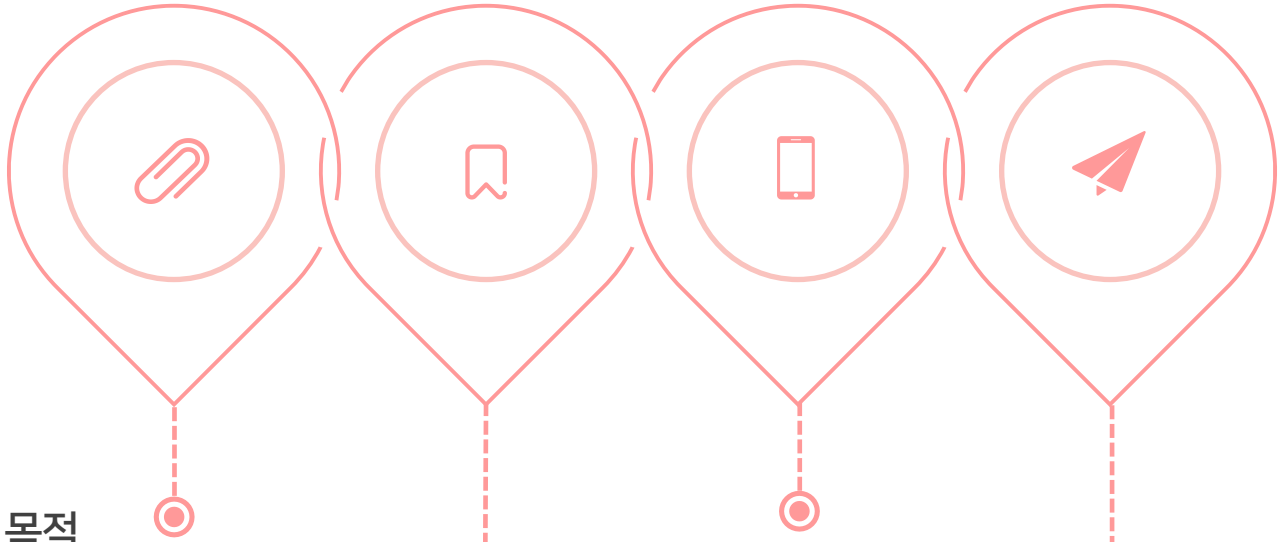
신용카드 이상거래 탐지

신보람 최다희 정다혜 박지원

목차



데이터 소개 및 분석 목적

데이터 소개 및 분석 목적

변수 설명

데이터 전처리

파생 변수 생성

변수 제거

스케일링

모델링

불균형 데이터 해결하기

① sampling / random forest

② OC-SVM

결과 및 한계점

분석 결과

한계점

1. 데이터 소개 및 분석 목적

데이터 소개 및 분석 목적

2013년 9월 유럽 카드 소유자들의 신용카드 거래 정보 (조사기간 : 2일)

Time

거래 시간(초)

V1-V28

PCA처리된 data

Amount

거래 금액

신용카드 이상 거래 여부를 판단

데이터 소개 및 분석 목적

Time	V1	V2	…	V28	Amount	Class
0	-1.3598071336738	-0.0727811733098497	…	-0.0210530534538215	149.62	0
0	1.3598071336738	0.0727811733098497	…	0.266150712059637	2.69	0
…	…	…	…	…	…	…
172788.0	-0.240440049680947	0.530482513118839	…	0.104532821478796	10	0
172792.0	-0.53341252200504	-0.189733337002305	…	0.0136489143320671	217	0

행(거래 건수) : 284,807개

변수 : 31개

정상 거래 : 284,315건 (99.8%)

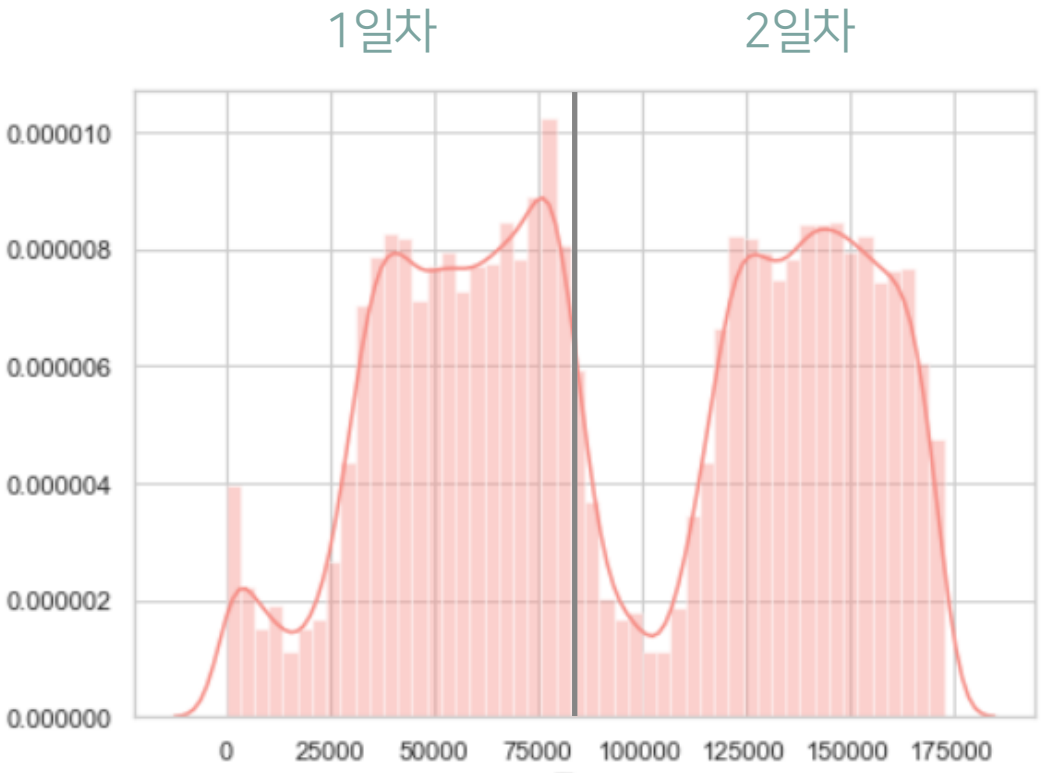
이상 거래 : 492건 (0.172%)

매우 불균형한 데이터

2. 데이터 전처리

데이터 전처리

Hour 변수 생성 및 Time 변수 제거



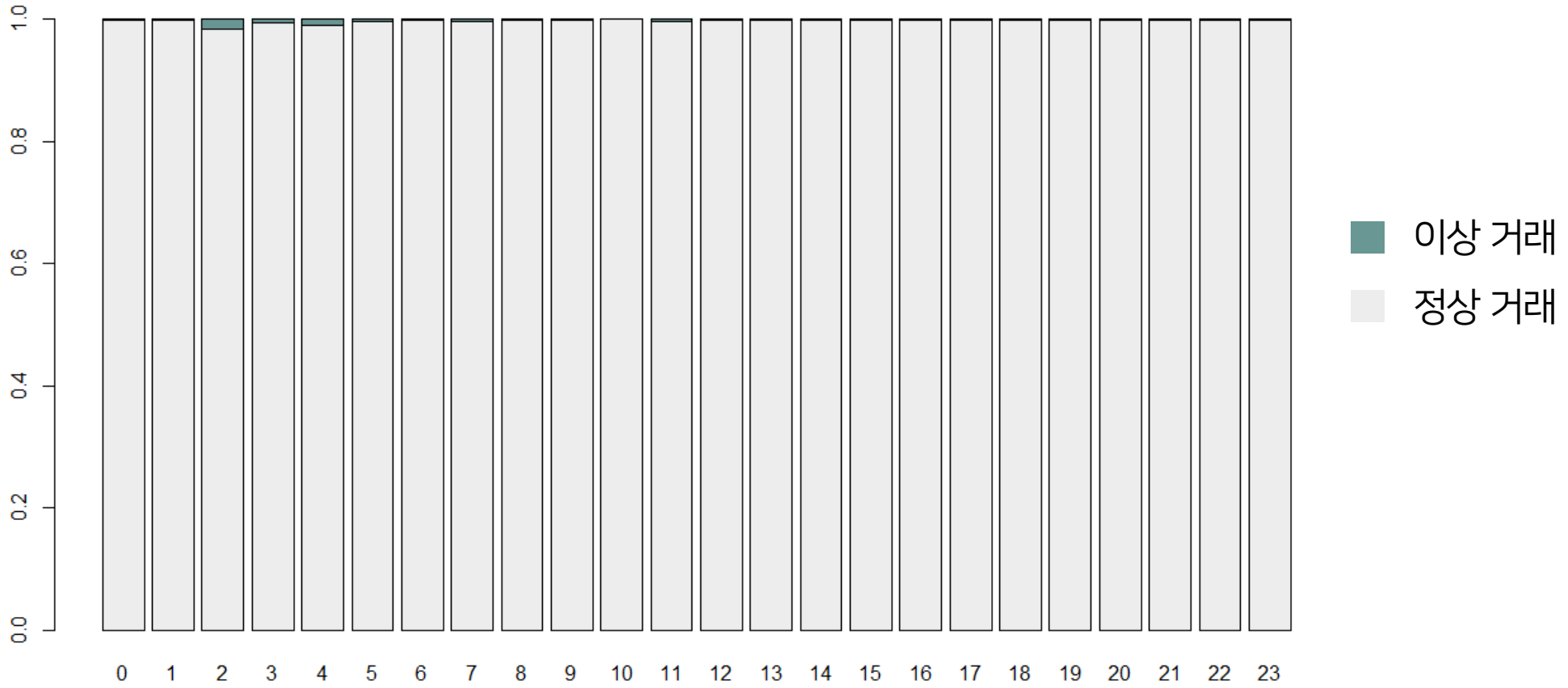
Time 변수를 통해 Hour 변수 생성

1일차	60×60×0(초) ≤ < 60×60×2(초)	0(h)
	⋮	
	60×60×23(초) ≤ < 60×60×24(초)	23(h)
2일차	60×60×24(초) ≤ < 60×60×25(초)	0(h)
	⋮	
	60×60×47(초) ≤ < 60×60×48(초)	23(h)

데이터 전처리

Hour 변수 생성 및 Time 변수 제거

[생성된 Hour 변수 그래프]



데이터 전처리

V변수 제거

Class 분류에 미치는 영향이 적은 경우

① 각 주성분들의 표준편차 제곱 값 < 0.7

② Class에 따른 histogram 유사

③ Class에 따른 box plot 유사

3가지 조건을 모두 만족하는 V변수를

Class 분류에 영향을 주지 않는다고 판단

V20 - V28 변수 제거

데이터 전처리

V변수 제거

① 각 주성분들의 표준편차 제곱 값이 0.7 미만인 변수

변수명	표준편차 제곱 값	변수명	표준편차 제곱 값
V1	3.836476	V15	0.837800
V2	2.726810	V16	0.767816
V3	2.299021	V17	0.721371
V4	2.004677	V18	0.702537
V5	1.905074	V19	0.662660
V6	1.774940	V20	0.594323
V7	1.530395	V21	0.539524
V8	1.426474	V22	0.526641
V9	1.206988	V23	0.389950
V10	1.185590	V24	0.366807
V11	1.041851	V25	0.271730
V12	0.998400	V26	0.232542
V13	0.990567	V27	0.162919
V14	0.837800	V28	0.108955

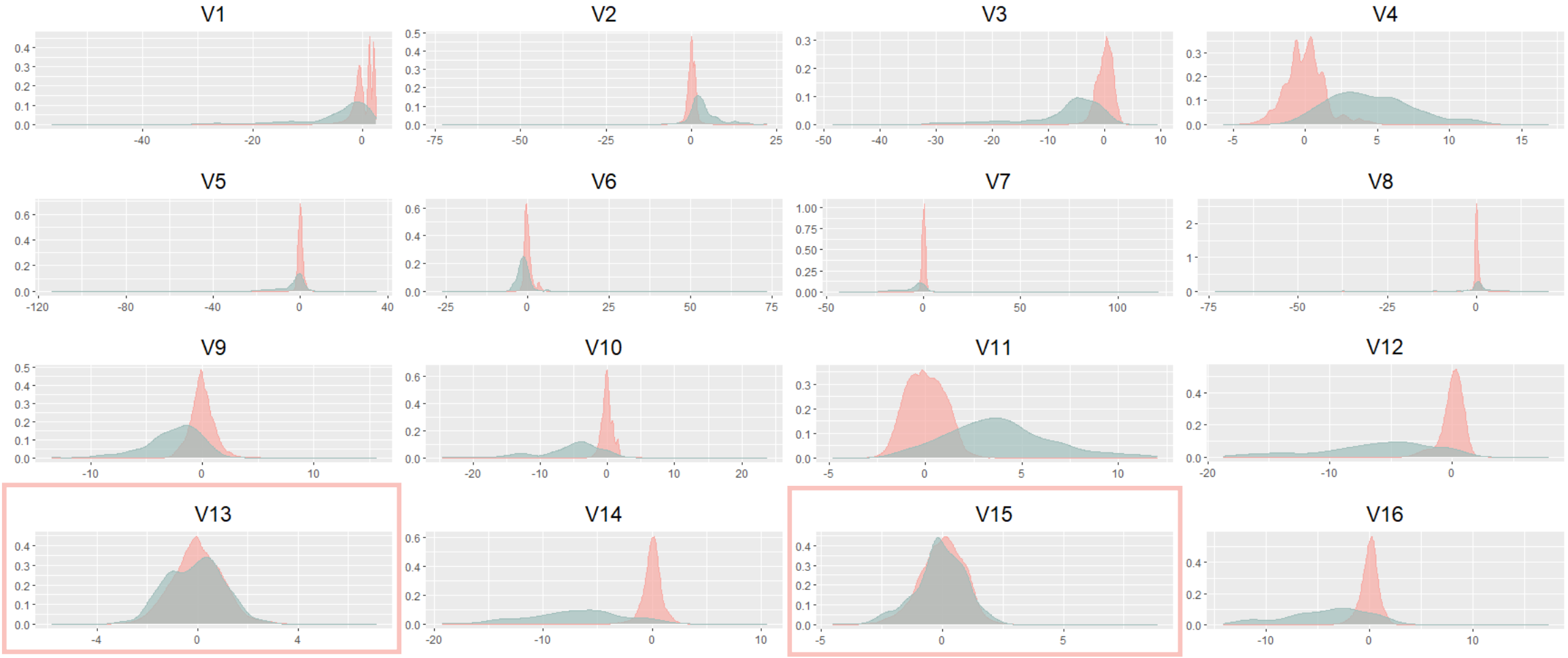
V19-V28

데이터 전처리

정상 거래
이상 거래

V변수 제거

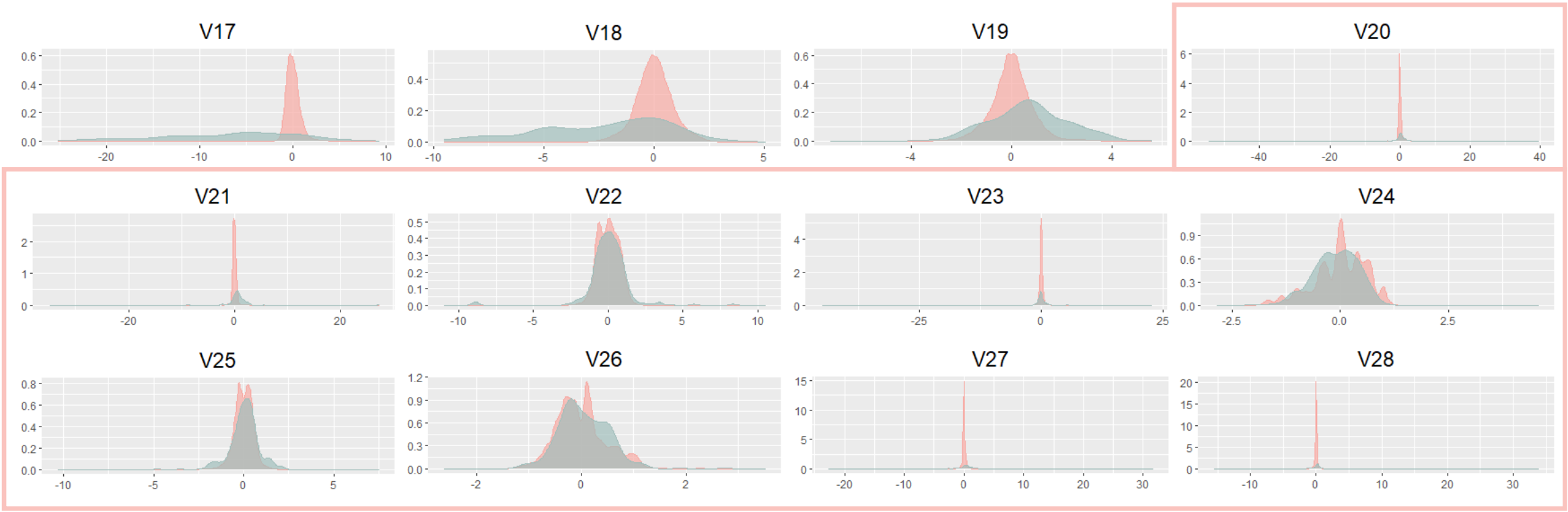
② Class에 따른 histogram 유사한 변수 : V13, V15



데이터 전처리

정상 거래
이상 거래

V변수 제거 ② Class에 따른 histogram 유사한 변수 : V20-V28

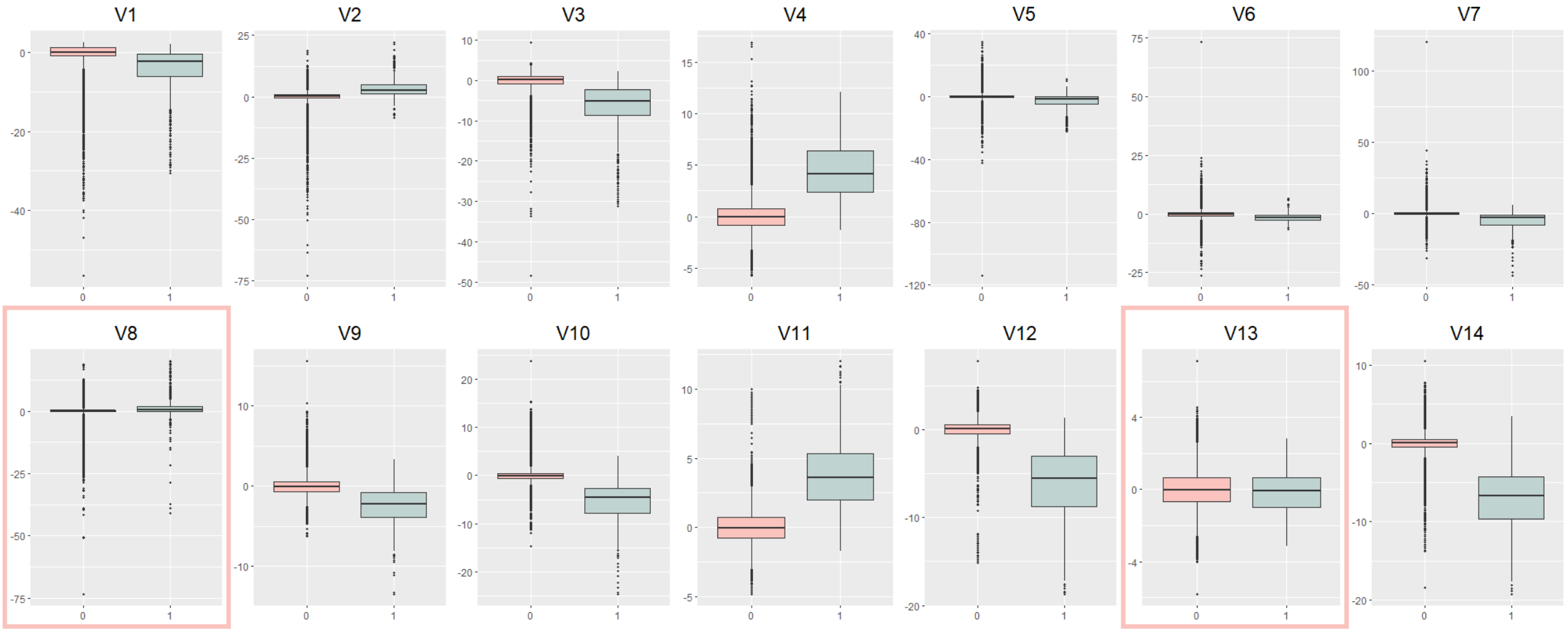


데이터 전처리

정상 거래
이상 거래

V변수 제거

③ Class에 따른 box plot 유사한 변수 : V8, V13

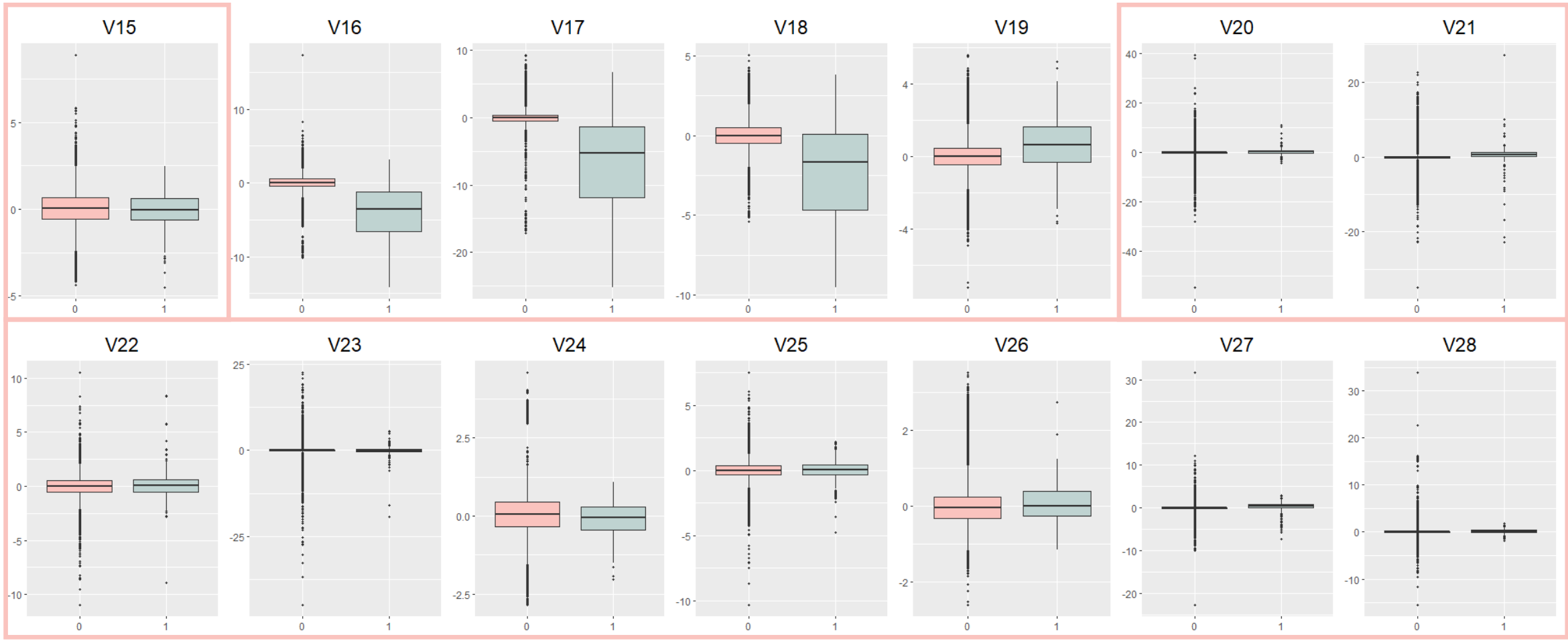


데이터 전처리

정상 거래
이상 거래

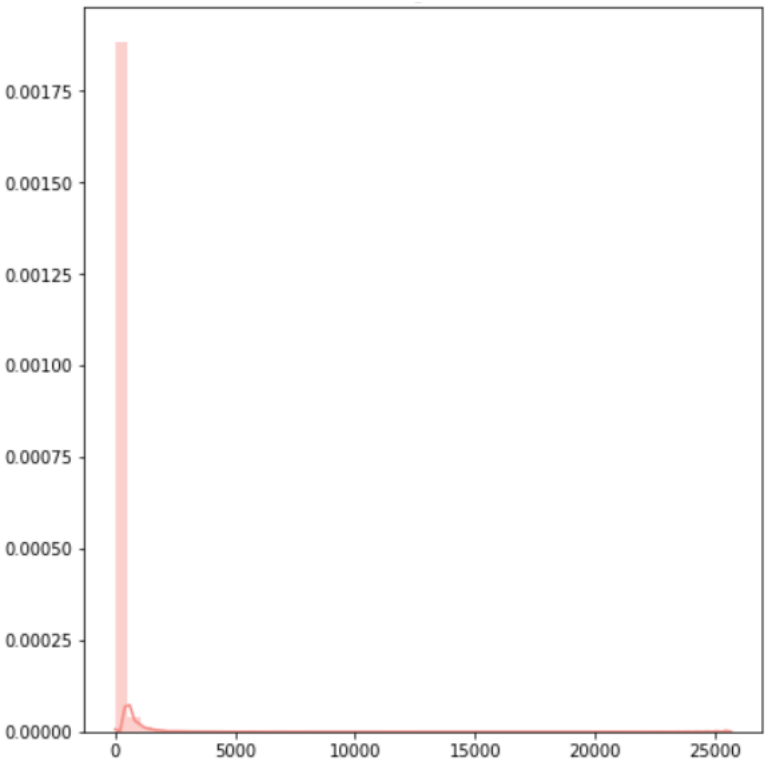
V변수 제거

③ Class에 따른 box plot 유사한 변수 : V15, V20–V28



데이터 전처리

Robust scaling – Amount 변수



Amount

범위가 0~25691로 매우 넓은 분포

범위를 줄이기 위해 Robust scaling 진행

데이터 전처리

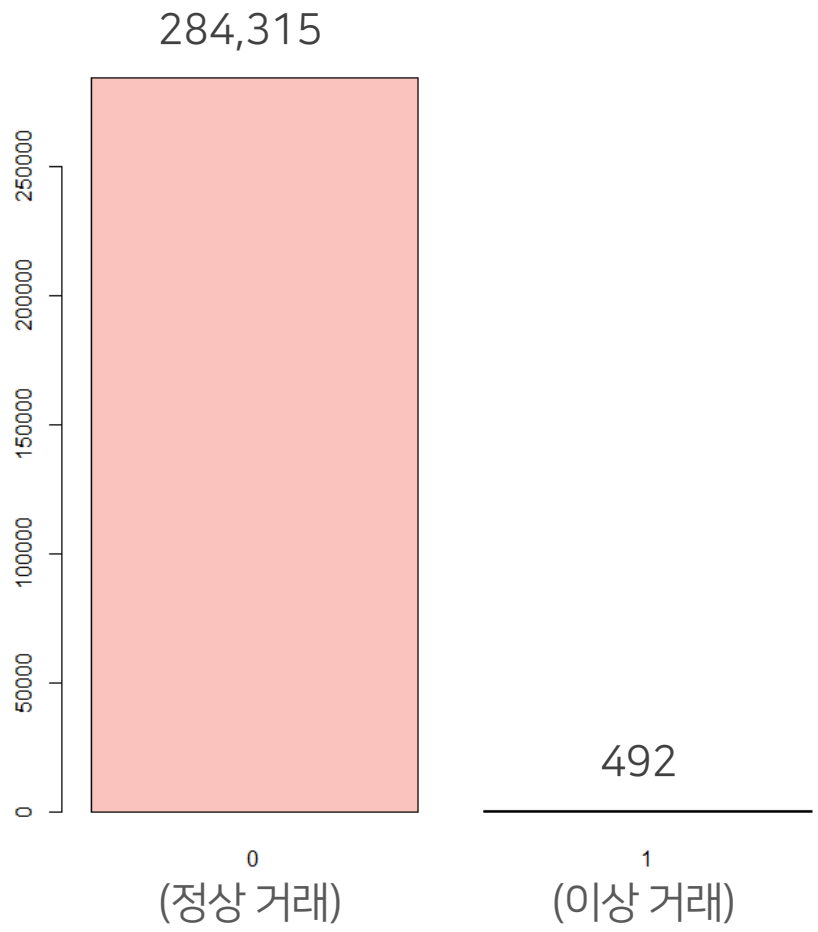
Robust scaling – Amount 변수

	스케일링 전	스케일링 후
Max	25691.16	358.6832
Q3	77.165	0.7708
Q2	22	0
Q1	5.6	-0.2292
Min	0	-0.3074
Std	250.1201	3.4950
mean	88.3496	0.9271

3. 모델링

모델링

불균형 데이터 해결하기



종속변수인 Class의 분포가 매우 불균형

Sampling

- SMOTE TOMEK
- SMOTE ENN

OCSVM
(One Class SVM)

모델링

① Sampling

Train set과 Test set을 7:3 비율로 split

	정상 거래	이상 거래	총 obs. 개수
Train set	199,020	344	199,364
Test set	85,295	148	85,443

모델링

① Sampling

Train data에 대해서 sampling 진행

SMOTE ENN

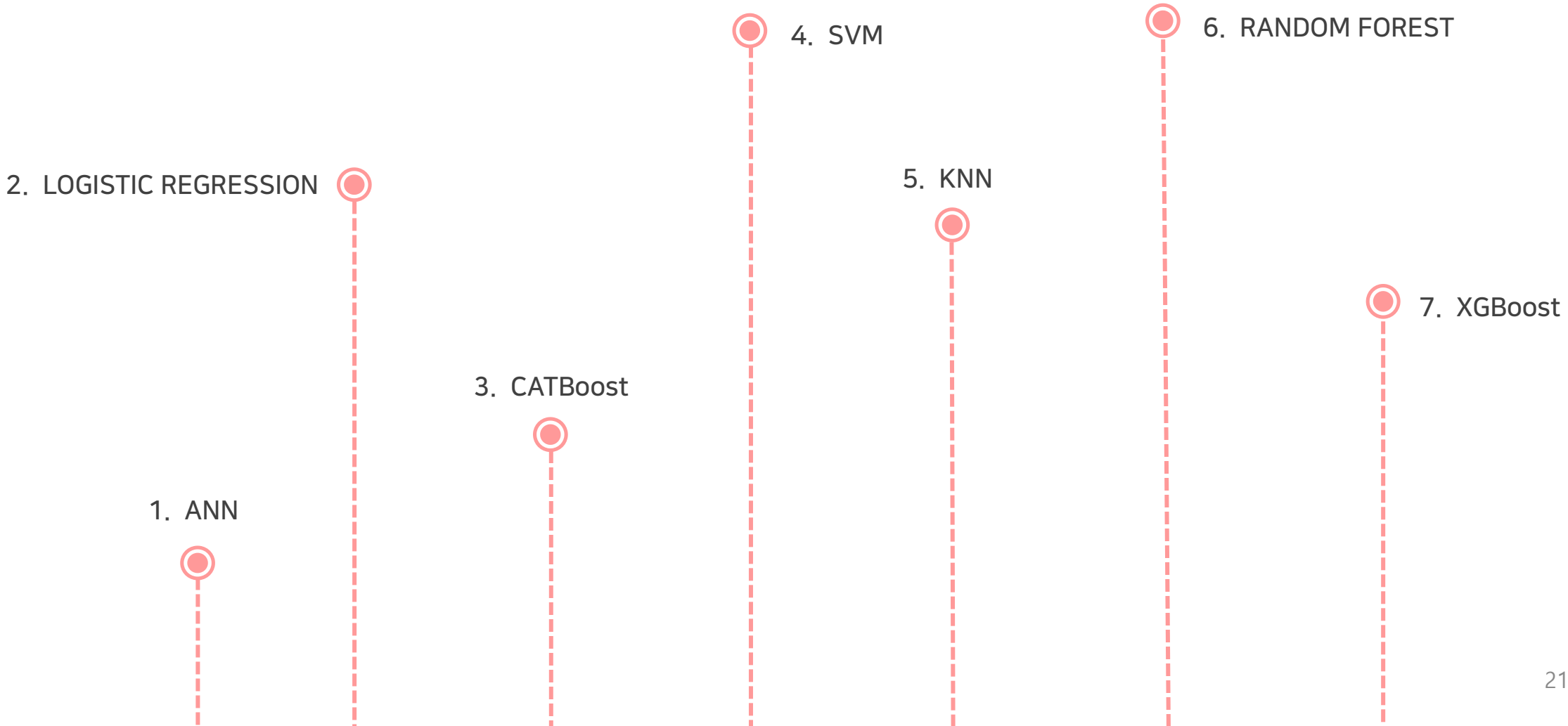
다수의 클래스 데이터 중 가장 가까운 K개의 데이터들이 소수의 클래스 데이터를 포함하면 해당 다수의 클래스 데이터를 삭제하는 방법
+ SMOTE

SMOTE TOMEK

가까이 붙어 있는 서로 다른 클래스의 데이터 중 다수 클래스에 속하는 데이터를 삭제하는 방법
+ SMOTE

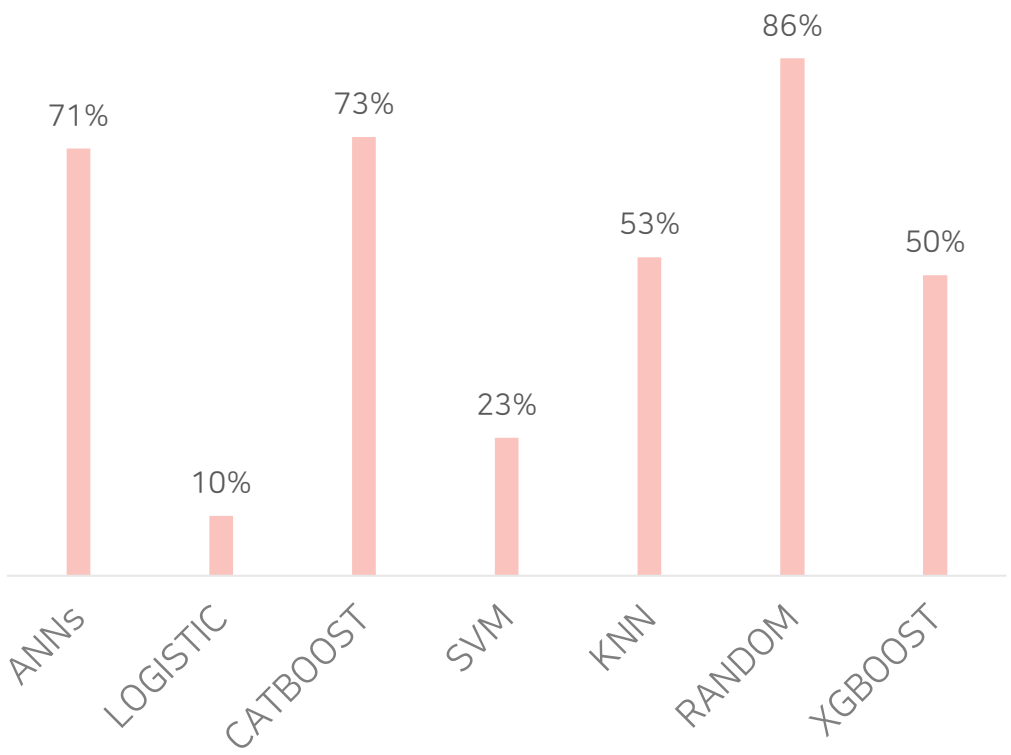
모델링

① Sampling

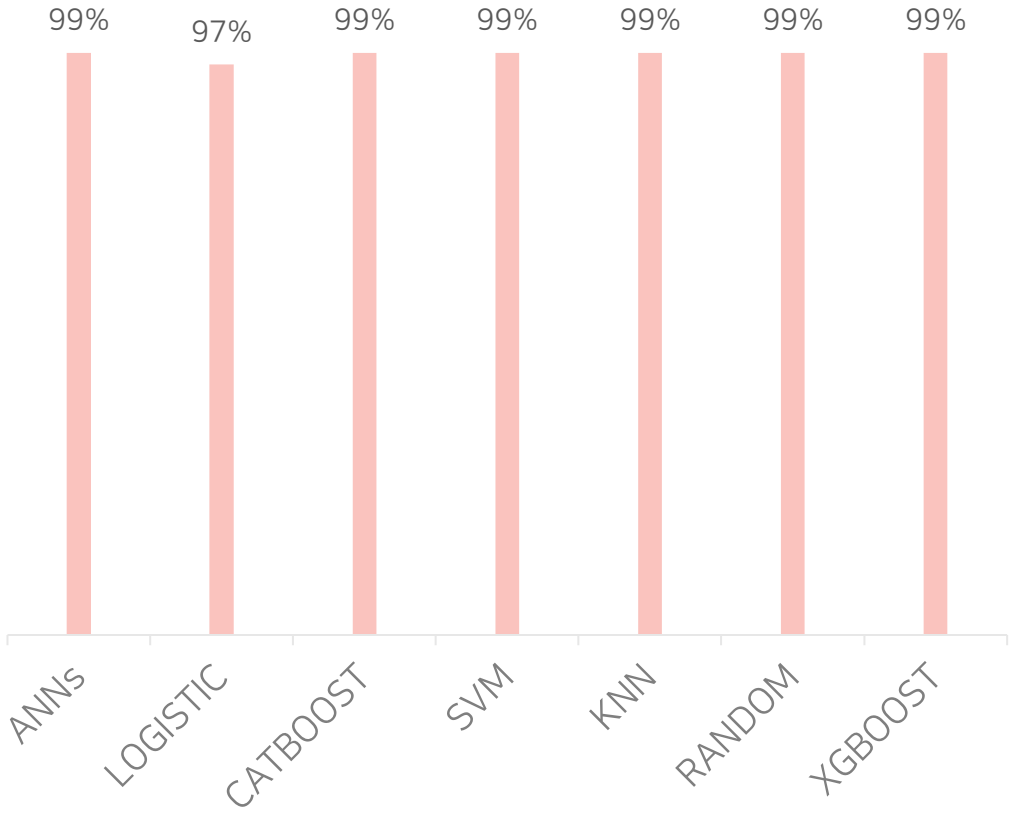


모델링

① Sampling - SMOTE ENN



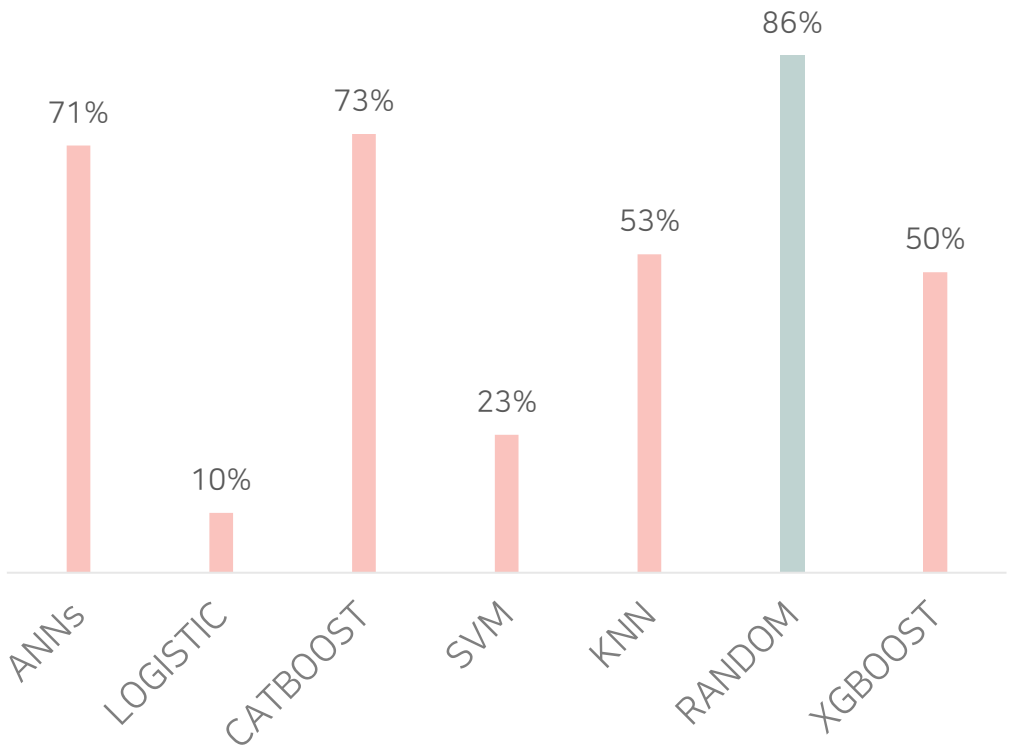
F1-score



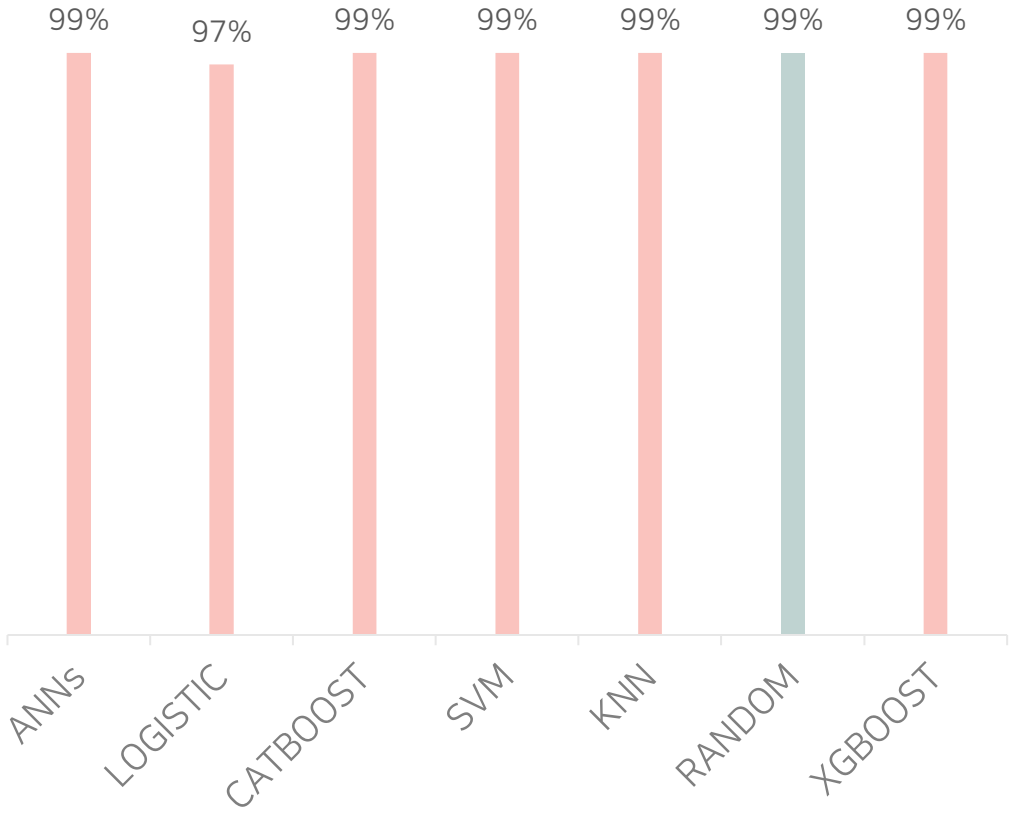
ACCURACY

모델링

① Sampling - SMOTE ENN



F1-score



ACCURACY

모델링

① Sampling - SMOTE ENN

T=0 F=1

T P	F N
F P	T N

혼동행렬

predicted

true

82,526	81
16	120

ANN

83,063	2,232
13	135

LOGISTIC

85,235	60
26	122

CATBOOST

84,525	771
27	120

SVM

85,095	200
21	127

KNN

85,272	23
18	130

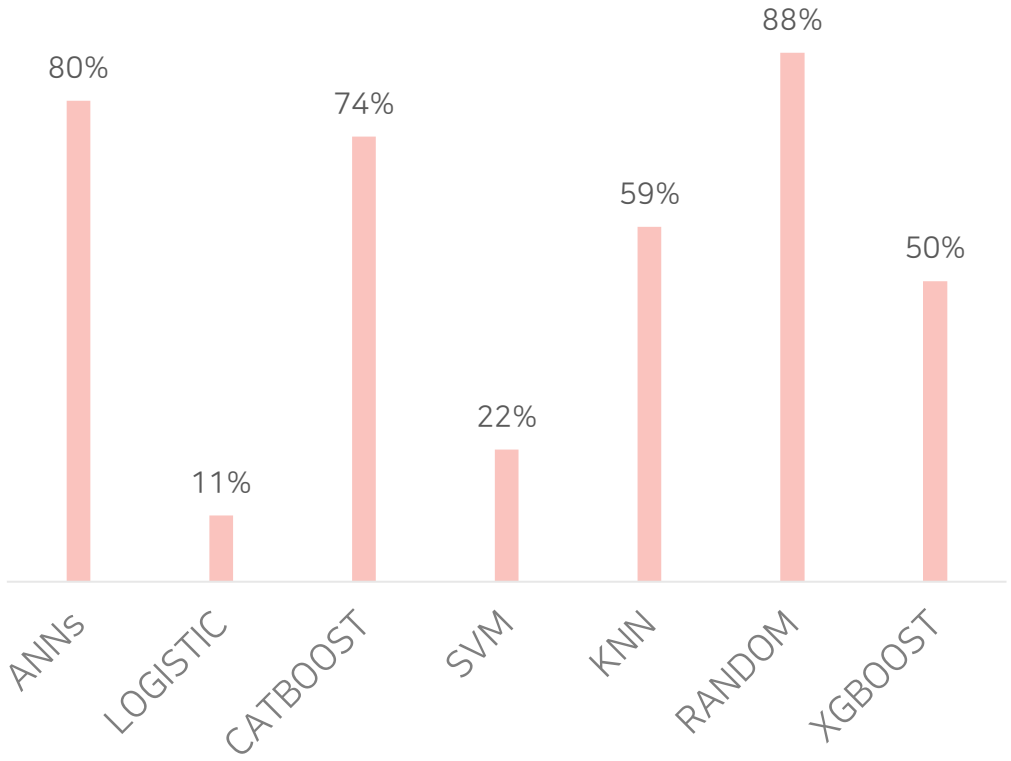
RANDOMFOREST

85,033	262
11	137

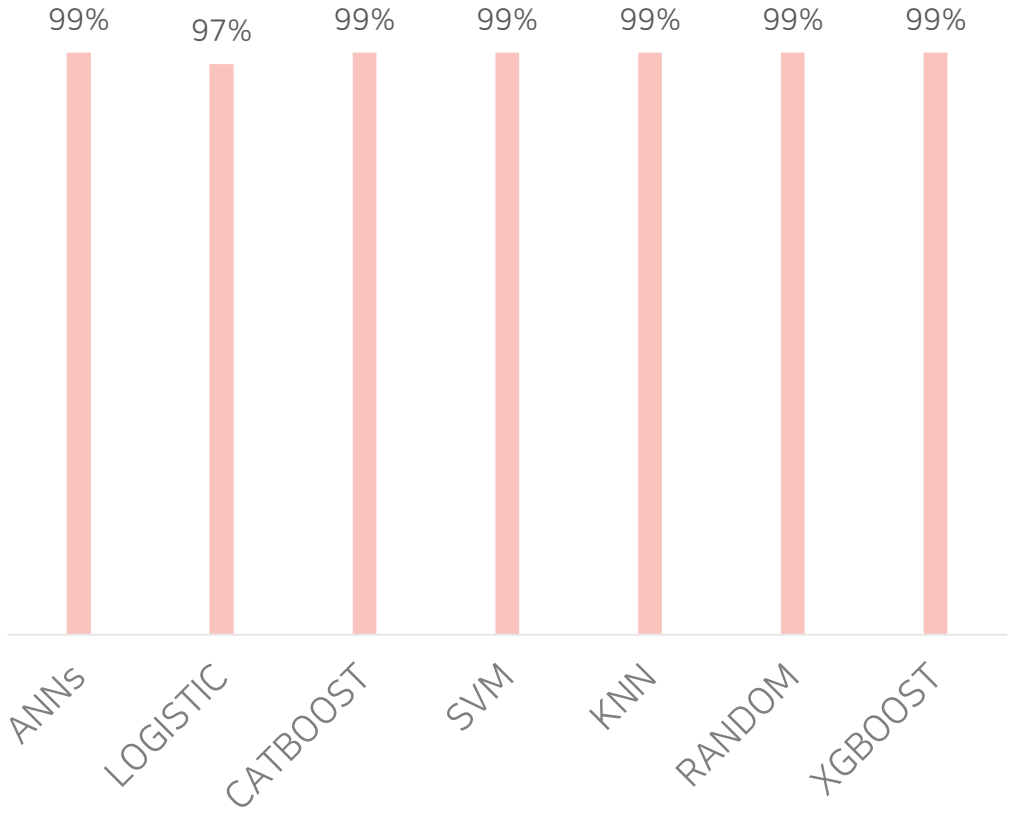
XGBOOST

모델링

① Sampling - SMOTE TOMEK



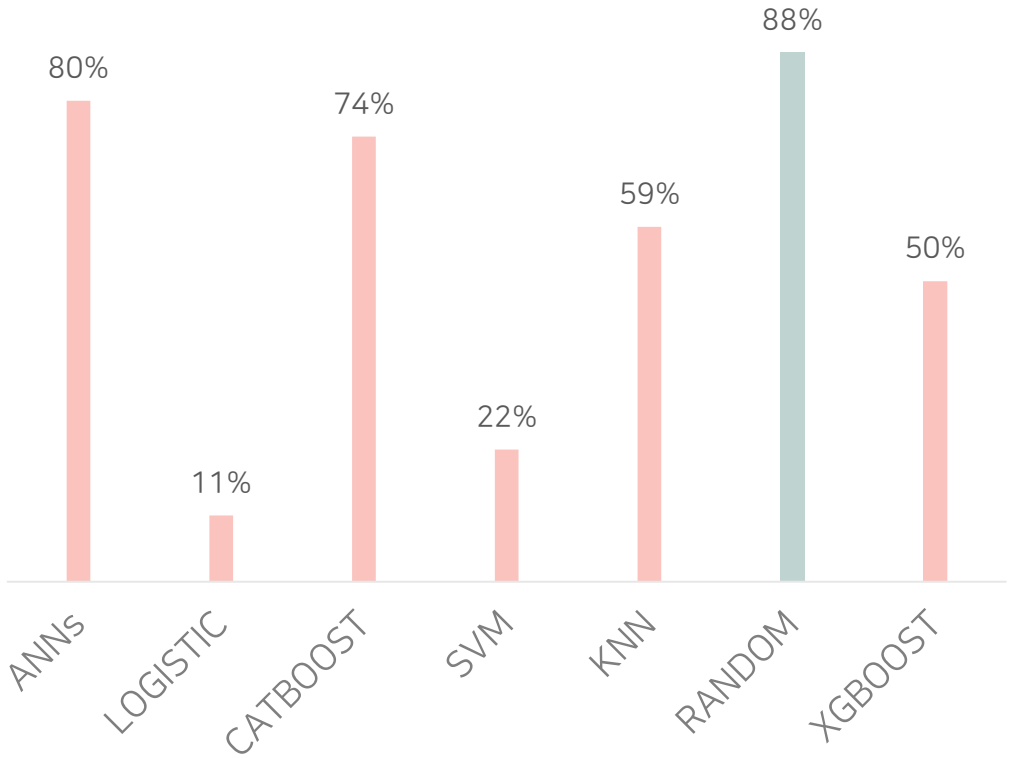
F1-score



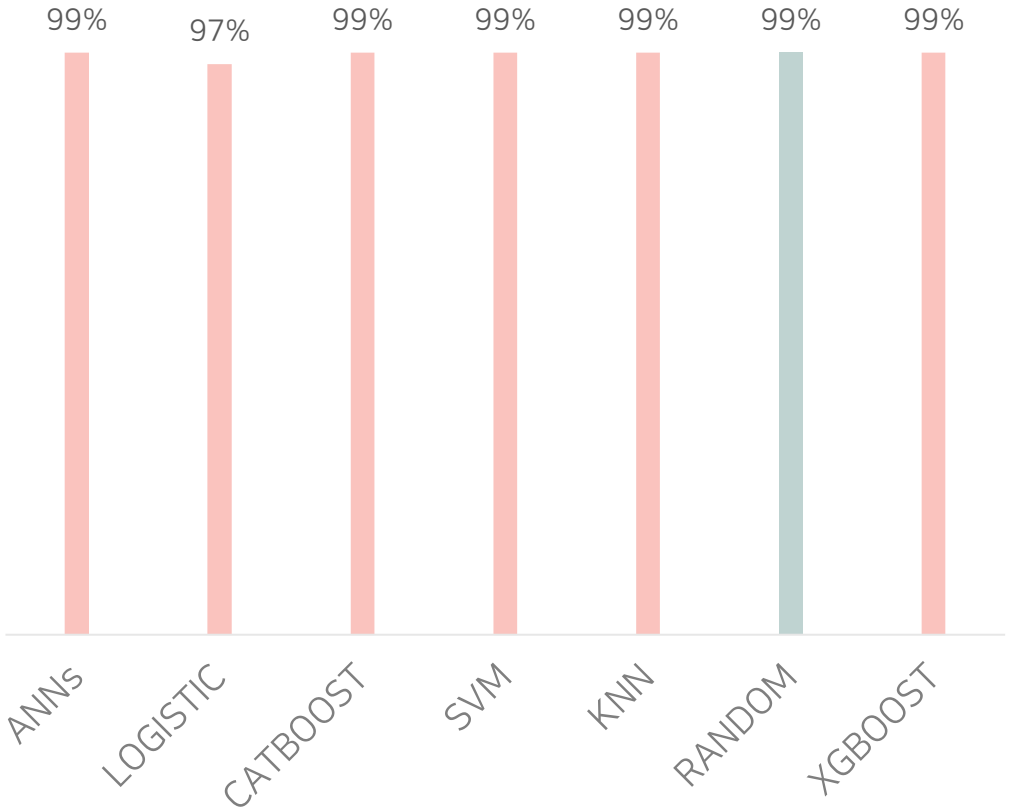
ACCURACY

모델링

① Sampling - SMOTE TOMEK



F1-score



ACCURACY

모델링

① Sampling - SMOTE TOMEK

T=0 F=1

T P	F N
F P	T N

혼동행렬

predicted

true

85,275	32
22	114

ANN

83,200	2,095
14	134

LOGISTIC

85,235	60
26	122

CATBOOST

84,511	785
27	120

SVM

85,143	152
23	125

KNN

85,279	16
19	129

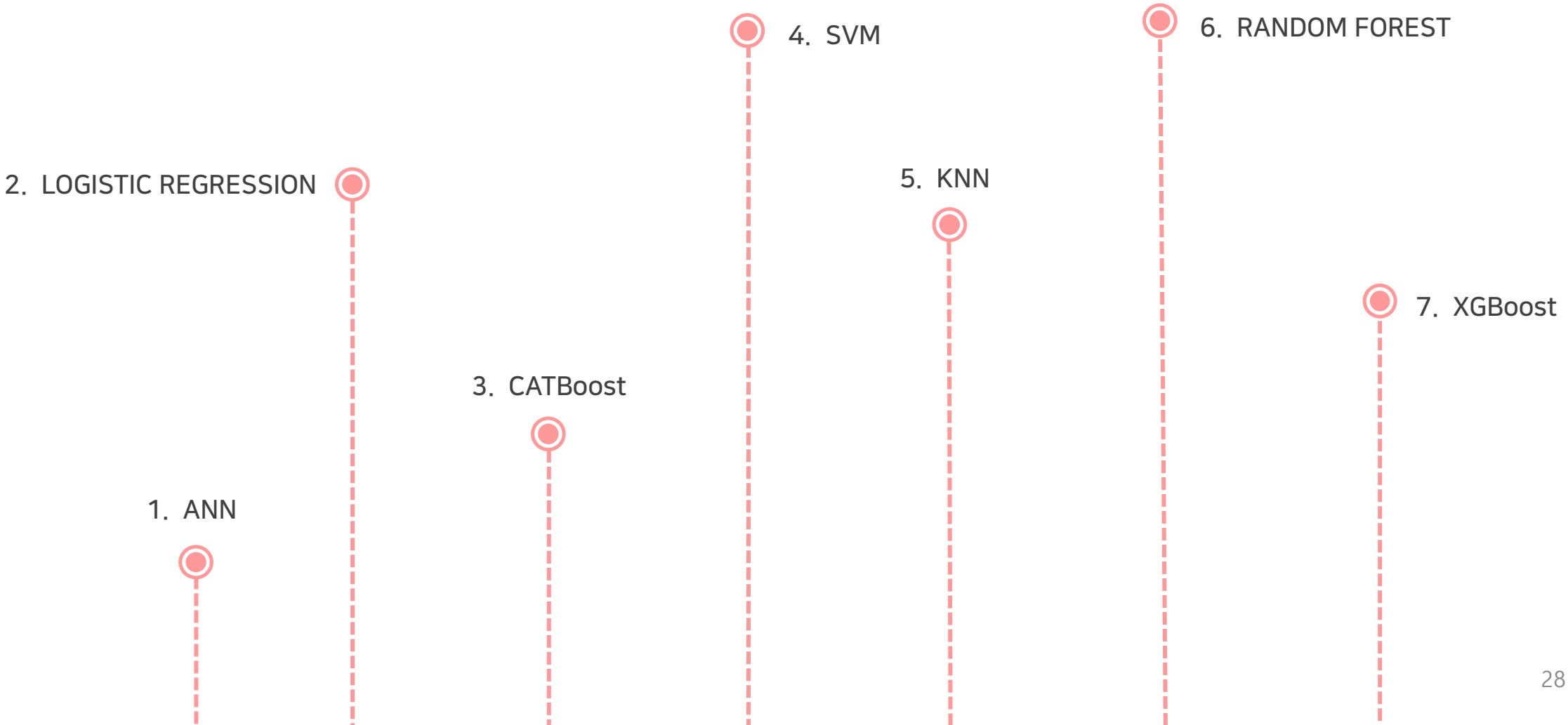
RANDOMFOREST

85,044	251
14	134

XGBOOST

모델링

① Sampling



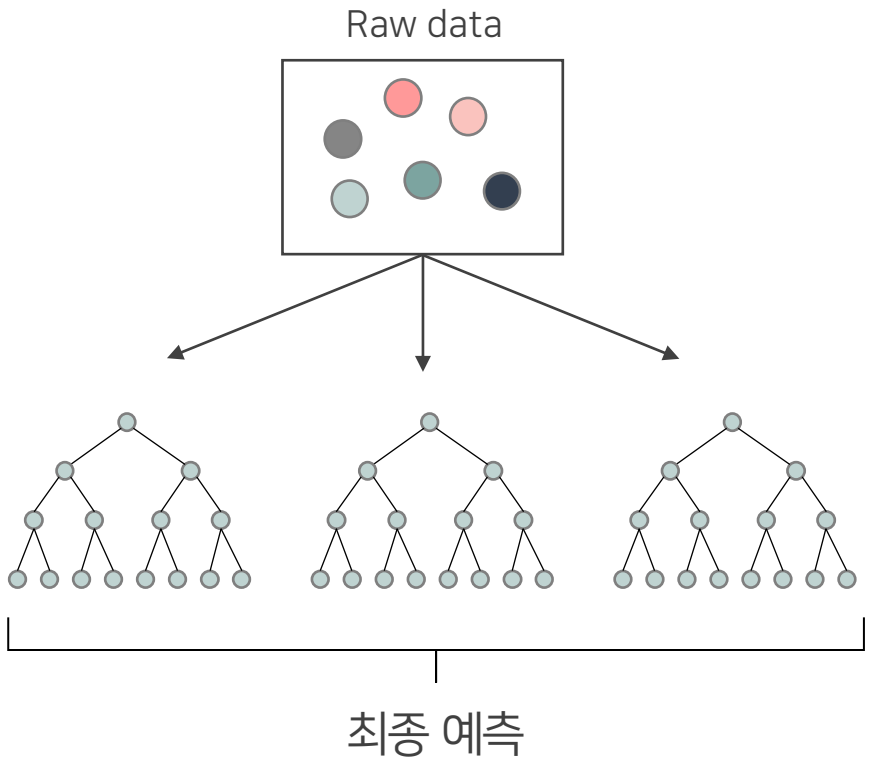
모델링

① Sampling

1. ANN
2. LOGISTIC REGRESSION
3. CATBoost
4. SVM
5. KNN
6. RANDOM FOREST
7. XGBoost

모델링

Random Forest (랜덤 포레스트)



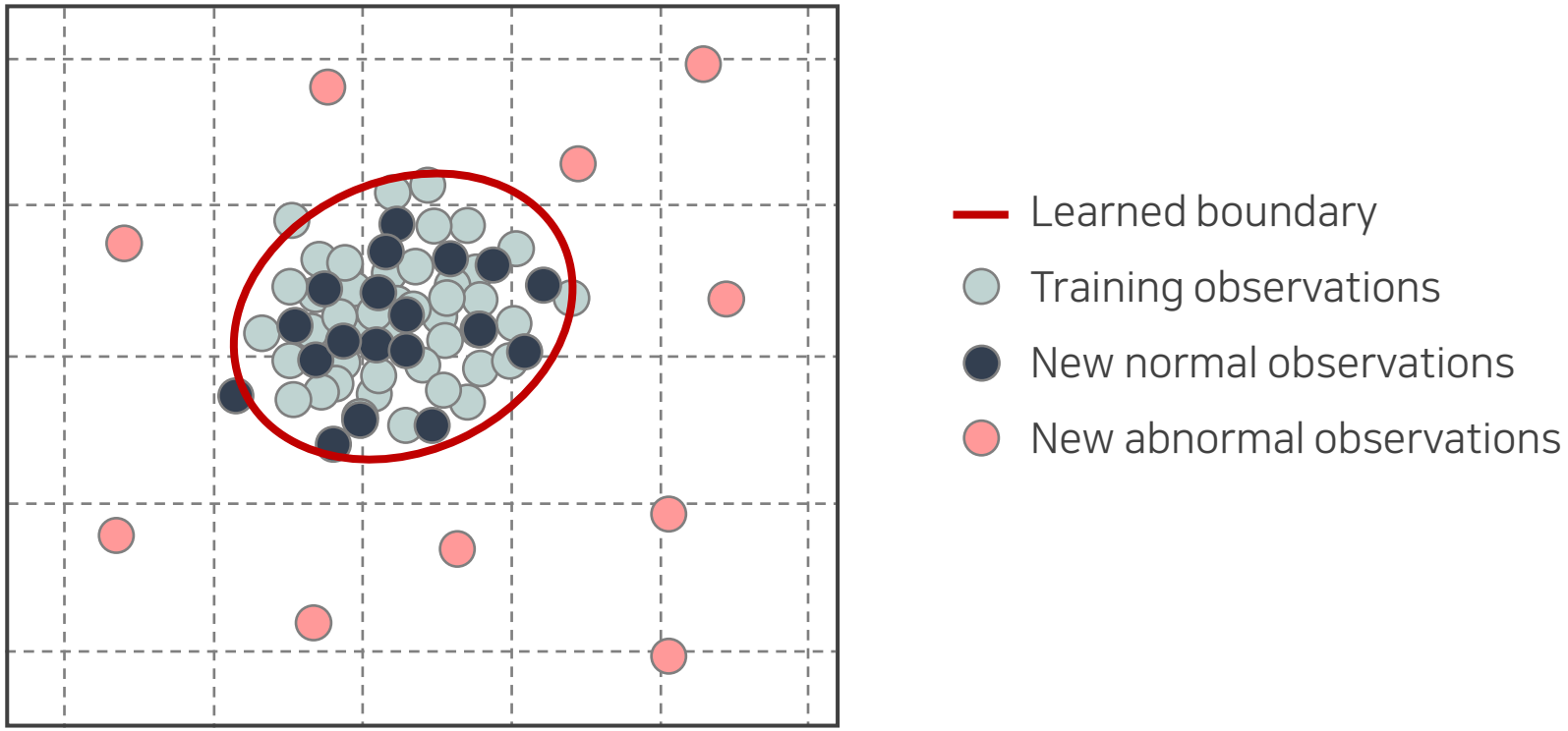
다수의 결정 트리들을 학습하는

앙상블(Ensemble) 방법

- 단일 트리의 단점을 보안
- 예측의 변동성 감소
- 과적합 방지

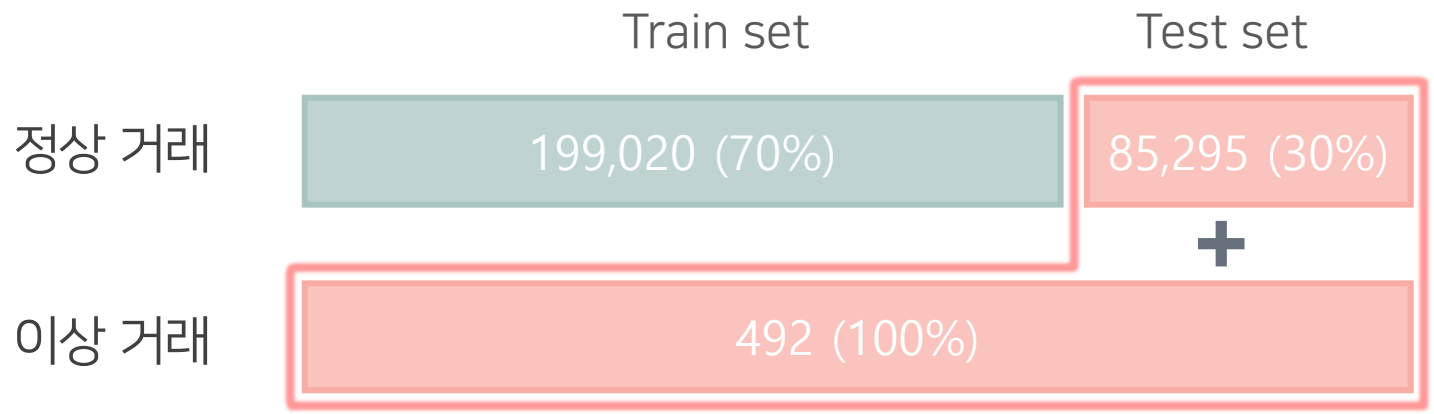
② OCSVM (One Class SVM)

이진 분류에서 하나의 Class만 학습시켜 불균형 데이터를 예측하는 비지도 학습 알고리즘



모델링

② OCSVM (One Class SVM)



정상 거래 데이터만 이용하여 모델 학습시킨 후,

테스트 데이터로 모델 성능 평가

4. 결과 및 한계점

결과 및 한계점

분석 결과

Random Forest - SMOTE TOMEK 기법 사용

예측 값

실제 값	정상 거래	이상 거래
정상 거래	85,279	16
이상 거래	19	129

- 정확도 : 0.99
- 재현율 : 0.87
- 정밀도 : 0.89
- F1 score : 0.88

결과 및 한계점

분석 결과

OC-SVM (One Class SVM)

예측 값

실제 값	정상 거래	이상 거래
정상 거래	78,449	6,846
이상 거래	87	405

- 정확도 : 0.92
- 재현율 : 0.92
- 정밀도 : 0.99
- F1 score : 0.96

결과 및 한계점

한계점

1. 종속변수의 범주 불균형이 너무 심하여 분석이 어려웠다. 불균형을 해결하기 위해 sampling을 하고 모델링을 진행했지만 과적합이 발생하였다.

2. 독립변수의 경우 이미 대부분 PCA 처리된 데이터가 주어져 변수에 대한 정보가 부족했다.

감사합니다